PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION

(PCT Rule 61.2)

To:

Commissioner
US Department of Commerce
United States Patent and Trademark Office, PCT
2011 South Clark Place Room CP2/5C24
Arlington, VA 22202
ETATS-UNIS D'AMERIQUE

in its capacity as elected Office

Date of mailing (day/month/year) 05 July 2001 (05.07.01)	
International application No. PCT/DE00/01791	**Applicant's or agent's file reference** 99P2291P
International filing date (day/month/year) 31 May 2000 (31.05.00)	**Priority date** (day/month/year) 20 July 1999 (20.07.99)
Applicant BAYER, Thomas	

1. The designated Office is hereby notified of its election made:

[X] in the demand filed with the International Preliminary Examining Authority on:

15 February 2001 (15.02.01)

[ ] in a notice effecting later election filed with the International Bureau on:

2. The election [X] was

[ ] was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland**	Authorized officer H. Zhou
Facsimile No.: (41-22) 740.14.35	Telephone No.: (41-22) 338.83.38

THIS PAGE BLANK (USPTO)

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
25. Januar 2001 (25.01.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/06451 A1

(51) Internationale Patentklassifikation[7]: G06K 9/72, 9/62

(21) Internationales Aktenzeichen: PCT/DE00/01791

(22) Internationales Anmeldedatum:
31. Mai 2000 (31.05.2000)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
199 33 984.8 20. Juli 1999 (20.07.1999) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, D-80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **BAYER, Thomas** [DE/DE]; Höriblick 10, D-78315 Radolfzell (DE).

(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, D-80506 München (DE).

(81) Bestimmungsstaaten *(national)*: AU, CA, JP, NZ, US.

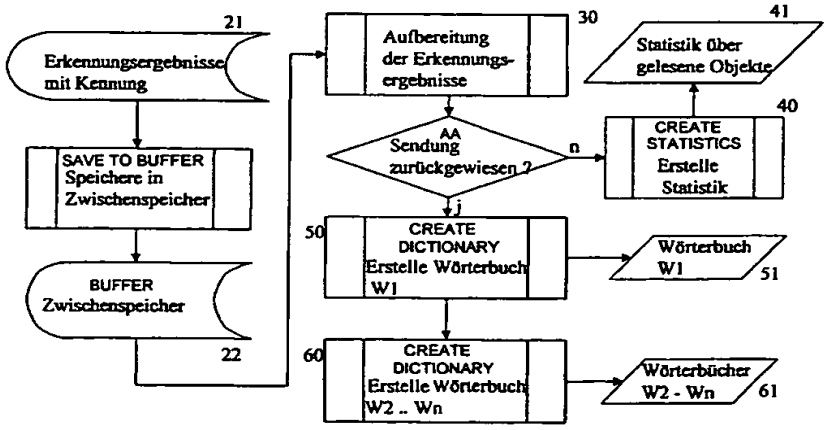
(84) Bestimmungsstaaten *(regional)*: europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht:
— *Mit internationalem Recherchenbericht.*
— *Vor Ablauf der für Änderungen der Ansprüche geltenden Frist; Veröffentlichung wird wiederholt, falls Änderungen eintreffen.*

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD FOR CREATING AND/OR UPDATING DICTIONARIES FOR AUTOMATICALLY READING ADDRESSES

(54) Bezeichnung: VERFAHREN ZUR BILDUNG UND/ODER AKTUALISIERUNG VON WÖRTERBÜCHERN ZUM AUTOMATISCHEN ADRESSLESEN



21...IDENTIFICATION RESULTS WITH IDENTIFIER
30...PREPARATION OF IDENTIFICATION RESULTS
AA...TRANSMISSION REJECTED?
41...STATISTICS CONCERNING READ OBJECTS
51...DICTIONARY
61...DICTIONARIES

(57) Abstract: According to the invention, the read results of an agreed number of transmission images obtained by the OCR reader, divided into clearly legible and rejected read results are buffered. Subsequently, classes of words or related word groups of the buffered and rejected read results are formed, each consisting of n address words, n = 1, 2, ..., a, with word spacing m, m = 0, 1, ..., b. The respective words in said classes do not fall below a specific degree of similarity among themselves, in relation to one specific n and m value. At least the representatives of those classes whose frequency exceeds a defined value are included in the dictionary or dictionaries of the corresponding address zones.

[Fortsetzung auf der nächsten Seite]

WO 01/06451 A1

Z Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

(57) Zusammenfassung: Es werden die vom OCR-Leser erzielten Leseergebnisse einer vereinbarten Anzahl von Sendungsbildern, unterteilt in eindeutig gelesene und zurückgewiesene Leseergebnisse zwischengespeichert. Dann werden Klassen von Wörtern oder zusammengehörenden Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse, bestehend jeweils aus n Adreßwörtern, n = 1, 2, ..., a, mit den Wortabständen m, m = 0, 1, ..., b, gebildet, die bezogen auf jeweils einen bestimmten n- und m-Wert untereinander ein bestimmtes Ähnlichkeitsmaß nicht unterschreiten. Mindestens Repräsentanten derjenigen Klassen, deren Häufigkeit einen festgelegten Wert überschreiten, werden in das oder die Wörterbücher der zugeordneten Adreßbereiche aufgenommen.

Beschreibung

Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adreßlesen

Die Erfindung betrifft ein Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum Adreßlesen.

Adreßlesesysteme benötigen Informationen über Inhalt und Syntax von Adressen, um die erforderlichen Informationen wie Stadt, Postleitzahl, Vorname und Nachname, etc. extrahieren zu können. Der zulässige Inhalt einzelner Adreßelemente wird mit einem Wörterbuch (Liste von zulässigen Zeichenketten) beschrieben, das nach dem Stand der Technik aus vorliegenden Informationsquellen aufgebaut wird, wie z.B. aus einem postalischen Wörterbuch oder aus einer Mitarbeiterliste einer Firma. Die Anwendungsdomäne ändert sich jedoch mit der Zeit, so daß das zu Beginn erstellte Wörterbuch nicht mehr alle vorkommenden Inhalte vollständig umfaßt. Vor allem bei der Anwendung eines Lesesystems zur innerbetrieblichen Postverteilung ist die Änderung des Wortvorrats beträchtlich: Mitarbeiter verlassen die Firma, neue Mitarbeiter kommen hinzu, Mitarbeiter wechseln die Abteilung oder Nachnamen ändern sich aufgrund von Heirat, etc. So fehlen im Wörterbuch Einträge und es gibt Einträge, die nicht mehr gültig sind. Je deutlicher der aktuell verwendete Wortvorrat vom Lexikon abweicht, desto mehr sinkt die Erkennungsleistung des Lesesystems.

Diese Änderungen mußten bisher in bestimmten Zeitabständen manuell in die Wörterbücher übertragen werden, so daß die geschilderten Nachteile auftraten.

Aufgabe der Erfindung ist es, ein Wörterbuch zum Adreßlesen automatisch zu bilden und/oder automatisch zu aktualisieren.

Erfindungsgemäß wird die Aufgabe durch die Merkmale des Anspruches 1 gelöst. Dabei wird von dem Gedanken ausgegangen, die Ergebnisse der aktuellen Leseprozesse zwischenzuspeichern,

auszuwerten und zum automatischen Aufbau oder zur Aktualisierung eines Wörterbuches zu nutzen. Beim Zwischenspeichern erfolgt eine Kennzeichnung, ob die jeweilige Adresse erfolgreich gelesen wurde oder ob sie zurückgewiesen wurde. Soll ein Wörterbuch neu erstellt werden oder sollen in das vorhandene Wörterbuch neue Adressaten aufgenommen werden, so werden die zurückgewiesenen Leseergebnisse herangezogen.

Die Wörterbücher können einzelne Wörter, z.B. Nachnamen und/oder zusammenhängende Wortgruppen mit n Wörtern, z.B. Vor- und Nachnamen oder Vor- und Nachnamen und Straßennamen enthalten, wobei die Wörter sowohl direkt nebeneinander (Abstand m=0) liegen als auch durch m Wörter beabstandet sein können.

Durch die Bildung von Klassen von Wörtern oder Wörtergruppen, die ein festgelegtes Mindestähnlichkeitsmaß zueinander besitzen, und die Aufnahme mindestens des Repräsentanten in das oder die Wörterbücher der zugeordneten Adreßbereiche, ist ein automatischer Aufbau eines Wörterbuches bzw. eine automatische Aktualisierung des Wörterbuches infolge neuer Adressaten oder von Änderungen bei den Adressaten möglich.

Vorteilhafte Ausgestaltungen der Erfindung sind in den Unteransprüchen beschrieben.

Zur Klassenbildung ist es vorteilhaft, eine Liste aller Wörter/Wortgruppen der zurückgewiesenen Leseergebnisse zu erstellen, die nach der Häufigkeit der Wörter/Wortgruppen sortiert ist. Dann wird, beginnend mit dem häufigsten Wort/Wortgruppe, das Ähnlichkeitsmaß mit allen übrigen Wörtern/Wortgruppen bestimmt und in eine Ähnlichkeitsliste eingetragen. Alle Wörter/Wortgruppen in der Ähnlichkeitsliste mit einem Ähnlichkeitsmaß über einer festgelegten Schwelle werden anschließend dem aktuellen Wort/Wortgruppe als Klasse zugeordnet. Danach werden die Wörter/Wortgruppen der gebildeten Klasse aus der Häufigkeitsliste entfernt.

Die Repräsentanten der jeweiligen Klasse von Wörtern oder Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse können durch die kürzesten oder häufigsten Wörter oder Wortgruppen gebildet werden.

Zur Erkennung von Adressen im Wörterbuch, die geändert oder entfernt werden müssen, ist es vorteilhaft, die eindeutig gelesenen Adressen statistisch auszuwerten. Tritt eine plötzliche Änderung der Häufigkeit der Wörter und/oder Wortgruppen über eine bestimmte Schwelle hinaus auf und dauert sie eine festgelegte Zeit an, so werden diese Wörter/Wortgruppen aus dem Wörterbuch entfernt.

Um zu vermeiden, daß irrelevante Wörter der Leseergebnisse in das Wörterbuch aufgenommen werden, können diese durch Vergleich mit in einer speziellen Datei für irrelevante Wörter gespeicherten Wörtern ermittelt werden.

Vorteilhaft in diesem Zusammenhang ist es auch, kurze Wörter ohne Abkürzungspunkt mit weniger als p Buchstaben als irrelevant nicht ins Wörterbuch aufzunehmen. Um die Adreßinterpretation mit Hilfe der Wörterbücher möglichst detailliert durchzuführen, ist es vorteilhaft, neben den Repräsentanten auch die Wörter und/oder Wortgruppen der dazugehörenden Klassen mit den Ähnlichkeitsmaßen und Häufigkeiten aufzunehmen.

In einer weiteren vorteilhaften Ausgestaltung können zusammengehörende Wortgruppen mit n Wörtern, die untereinander einen Abstand von m Wörtern haben, ermittelt werden, indem ausgehend vom jeweiligen, für das Wörterbuch ermittelten Einzelwort die Adressen mit Fenstern der Breite von n+m Wörtern durchsucht werden. Nachdem die weiteren n-1 Einzelwörter mit den Abständen von m Wörtern untereinander ermittelt wurden, erfolgt die Aufnahme dieser Wortgruppe mit ihren Häufigkeiten in das entsprechende Wörterbuch.

Vorteilhaft ist es auch, das Ähnlichkeitsmaß mit dem Levenshtein-Verfahren (siehe „A Method for the Correction of Garbled Words, Based on the Levenshtein Metric", K. Okuda, E. Tanaka, T. Kasai, IEEE Transactions on Computers, Vol. c-25, No. 2, February 1976) zu ermitteln.

Es kann auch vorteilhaft sein, die ermittelten Wörterbuchaktualisierungen an einem Videocodierplatz zu kategorisieren und bestätigen zu lassen oder die Neueintragungen ins Wörterbuch zusätzlich vor ihrer Übernahme in die entsprechende Kategorie mit den Inhalten einer Datei zu vergleichen, in der charakteristische allgemeingültige Namen oder wenigstens Zeichenstrings bezogen auf die jeweilige Kategorie (Vorname, Nachname, Abteilung) gespeichert sind.

Anschließend wird die Erfindung in einem Ausführungsbeispiel anhand der Zeichnung näher erläutert. Ziel hierbei ist, bisher unbekannte Nachnamen (n=1) oder Paare unbekannter Vor- und Nachnamen (n=2) oder Nach- und/oder Vor- und Nachnamen und Abteilungsnamen von Mitarbeitern einer Firma und/oder entsprechende nicht mehr gültige Namen bzw. Namenskombinationen zu ermitteln und Wörterbuchänderungen durchzuführen.

Dabei zeigen

FIG 1 eine Ablaufstruktur eines Monitorprozesses zur Überwachung und Steuerung der Aktualisierung des Wörterbuches

FIG 2 eine Ablaufstruktur zur Ermittlung und Kennzeichnung irrelevanter Wörter

FIG 3 eine Ablaufstruktur zur Ermittlung bisher unbekannter Einzelwörter (n=1) (Nachnamen)

FIG 4 eine Ablaufstruktur zur Ermittlung bisher unbekannter Wortgruppen, ausgehend von den Einzelwörtern

FIG 5 eine Ablaufstruktur zur Aktualisierung der Wörterbücher unter Berücksichtigung der Wortkategorien

Die Wortvorschläge werden aus den Erkennungsergebnissen automatisch generiert, die das Lesesystem im täglichen Betrieb für jedes Sendungsbild berechnet. Die Erkennungsergebnisse für jedes Sendungsbild umfassen unterschiedliche geometrische Objekte (Layoutobjekte), wie Textblöcke, Zeilen, Wörter und Zeichen, und deren Relationen untereinander, also, welche Zeilen zu welchem Textblock gehören, welche Wörter in welchen Zeilen liegen, etc. Für jedes Einzelzeichenbild erzeugt das Lesesystem eine Liste von möglichen Zeichenbedeutungen. Darüberhinaus berechnet das Lesesystem für jedes Layoutobjekt seine Lage im Sendungsbild und dessen geometrischen Ausmaße.
Zum Aktualisieren oder auch Lernen von Wörterbucheinträgen wird die Menge der bearbeiteten Sendungen in zwei Teilmengen getrennt, in die Menge der vom Lesesystem automatisch gelesenen (aber nicht notwendigerweise korrekt gelesenen) und die Menge der zurückgewiesenen Sendungen. Die Menge der automatisch gelesenen Sendungen dient zum Ermitteln von Wörterbucheinträgen, die nicht mehr gültig sind; aus der Menge der zurückgewiesenen Sendungen werden neue Wörterbucheinträge abgeleitet.

Das beispielhafte System besteht aus fünf Modulen: einen Monitorprozeß, einer Aufbereitung der Erkennungsergebnisse (Vorverarbeitung), zweier Wörterbuchgenerierungsverfahren und einem Vorschlagsadministrator.
Der Monitorprozeß gemäß FIG 1 überwacht und steuert das Wörterbuchlernen. Die Erkennungsergebnisse 21 für jedes Sendungsbild werden zusammen mit einer Kennung für „erfolgreich gelesen" oder „zurückgewiesen" vom Leser an den Monitor übergeben. Zusätzliche Informationen zur Sendungsart (Brief, Großbrief, Hauspostformular) und weitere Merkmale zu den einzelnen Objekten der Erkennungsergebnisse, wie ROI (Region of Interest), Zeilen- und Wort-Hypothesen, Zerlegungsalternativen und Schriftzeichen-Erkennungsergebnisse, können ebenfalls übergeben werden. Diese Erkennungsergebnisse werden im Monitor in einem Zwischenspeicher 22 gespeichert, bis eine genügend große Menge an Daten angefallen ist (z.B. nach 20.000 Sendungen oder nach einer Woche Betrieb).

Im einfachsten Fall wird lediglich die erste Alternative der Zeichenerkennungsergebnisse zusammen mit dem besten Segmentierpfad im Zwischenspeicher gespeichert. Beispielsweise könnte der Inhalt folgendermaßen aussehen:

<Erkennungsergebnisse>	<Kennung>
:...	
1017921 PMD 55	erkannt
MR. A1FRED C SCHMIDI	
EXCCU1LVE DIRCC1OR, 0PCRA1IONS	
DCVC1OPMENT	
MyComp, INC	
1 MyStreet	
MyCity, 12345	
PO11Y O/BRIEN	zurückgewiesen, nicht im Wörterbuch
MANAGER, COMMUNITY AFFAIRS	
MyComp INC	
1 MyStreet	
MyCity, 12345	
PO1LY OBRIEN	zurückgewiesen, nicht im Wörterbuch
MANAGER, COMMUNITY AFFAIRS	
MyComp, INC	
1 MyStreet	
MyCity, 12345	
MS ME1INDA DUCKSWORTH	erkannt
MyComp, INC	
MAI1 CODE 63-33	
1 MyStreet	
MyCity, 12345	

**********AUR0**MIXED AADC 460 zurückgewiesen, nicht im Wörterbuch

MIK0 SCHWARTZ
O AND T 26-OO
1 MyStreet
MyCity, 12345

.....

Liegen genügend Ergebnisse vor, werden die zurückgewiesenen Erkennungsergebnisse an eine Aufbereitungseinheit 30 übergeben und zu den beiden Teilprozessen zum Wörterbuchlernen für Einzelworte 50 und Wortgruppen 60 weitergeleitet. Im Falle einer erfolgreichen automatischen Erkennung werden die Ergebnisse an ein Statistikmodul übergeben 40. Wenn alle Sendungen verarbeitet worden sind, werden die Wort- und Wortgruppenlisten 41 des Statistikmoduls und der Wörterbuchlernprozesse 51, 61 gesammelt und mit einer geeigneten grafischen Oberfläche einer Bedienkraft zur Bestätigung vorgelegt.

In der Aufbereitungseinheit 30 werden irrelevante Wörter in den zurückgewiesenen Erkennungsergebnissen gekennzeichnet, die in der nachfolgenden Textanalyse nicht berücksichtigt werden (vgl. FIG 2). Diese Wörter werden als nicht relevant markiert aber nicht gelöscht, da die Wortnachbarschaft für den nachfolgenden Wörterbuchaufbau wichtig ist.

Im Verfahrensschritt Markieren irrelevanter Wörter 31, werden aus der Menge der Worthypothesen kurze Wörter markiert, beispielsweise diejenigen, die weniger als 4 Buchstaben lang sind und gleichzeitig keinen Abkürzungspunkt besitzen, und solche die zu weniger als 50% aus alphanumerischen Zeichen bestehen. Weiterhin werden solche Wörter markiert, die in einer speziellen Datei 32 enthalten sind, die für diese Anwendung häufige, aber irrelevante Wörter enthält. Bei der Anwendung der innerbetrieblichen Postverteilung können beispiels-

weise der Firmenname, Städtename, Straßenname, Postfachbezeichnung, etc., in diesem speziellen Lexikon enthalten sein. Die Ergebnisse der Aufbereitung werden in einen Zwischenspeicher 33 zurückgeschrieben.

Nach der Vorverarbeitung sehen die Ergebnisse folgendermaßen aus:

<title MR> <first-name ALFRED> <last-name SCHMID>
<role EXECUTIVE DIRECTOR OPERATIONS>

PO11Y O/BRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

PO1LY OBRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

<title MS> <first-name MELINDA> <last-name DUCKSWORTH>

<non-alpha **********AUR0**MIXED> AADC <short 460>
MIK0 SCHWARTZ
<short O> <short AND> <short T> 26-00
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

.....

Aus den aufbereiteten zurückgewiesenen Erkennungsergebnissen wird gemäß FIG 3 im ersten Schritt 52 eine Häufigkeitsliste FL 53 aller darin vorkommender Wörter erstellt, nach abstei-

gender Häufigkeit sortiert und in einen Zwischenspeicher abgelegt. Für obiges Beispiel könnte die Häufigkeitsliste FL 53 folgendermaßen aussehen:

...	
AFFAIRS	37
MANAGER	37
COMMUNITY	37
OBRIEN	20
O/BRIEN	17
SCHWARTZ	15
MIK0	12
POLLY	10
PO11Y	8
PAULA	8
PO1LY	5
MIKO	3
.....	

Aus dieser Liste wird schrittweise ein Wörterbuch W1 relevanter Wörter 51 aufgebaut. Zu jedem Wort in der Häufigkeitsliste FL 53 wird der Abstand d zu allen Wörtern in dieser Häufigkeitsliste bestimmt. Ein Verfahren zur Messung des Abstandes zwischen zwei Zeichenketten ist das Levenshtein-Verfahren, das den minimalen Abstand zweier Zeichenketten berechnet, bezogen auf 3 Kostenarten, auf Kosten einer Ersetzung eines Zeichens, einer Einfüge- und einer Löschoperation. Zur Berechnung von d können neben der Zeichenkette weitere Merkmale der Erkennungsergebnisse verwendet werden, beispielsweise die Zeichenalternativen, die Segmentieralternativen, etc.

Das erste Wort in der Häufigkeitsliste FL 53 (das aktuell häufigste) wird in das Wörterbuch W1 51 übernommen und aus der Häufigkeitsliste FL 53 gelöscht 54. Alle Wörter aus der Häufigkeitsliste FL 53 mit einem Abstand kleiner einer fest-

gelegten Schwelle th_d werden dem aktuellen Wort im Wörterbuch W1 51 mit ihrer Häufigkeit zugeordnet 55, 56. Gleichzeitig werden diese Wörter in der Häufigkeitsliste FL 53 gelöscht. Die Iteration endet, wenn die Häufigkeitsliste FL 53 leer ist. Damit werden Wortklassen gebildet, die untereinander einen Abstand d nicht überschreiten, bzw. ein entsprechendes Ähnlichkeitsmaß nicht unterschreiten.

Wenn alle Wörter verarbeitet sind, besteht das Wörterbuch W1 51 aus einer Menge von Wortklassen. Das kürzeste Wort einer Wortklasse wird als Repräsentant der Gruppe bezeichnet. Jede Wortklasse enthält Wörter, die einander ähnlich sind, mit den dazugehörigen Häufigkeiten und Abständen zum Klassenrepräsentanten. Die Repräsentanten der Wortklassen im Wörterbuch W1 51, und damit auch die Wortklassen, werden nach absteigender Häufigkeit sortiert 57. Die Häufigkeit einer Wortklasse setzt sich aus der Häufigkeit des Repräsentanten und der Häufigkeiten der Elemente der Wortklasse zusammen. Wortklassen, deren Häufigkeit eine gewisse Schwelle unterschreiten, werden aus dem Wörterbuch W1 51 gelöscht. Aus obiger Liste wird folglich folgendes Wörterbuch W1 51 gebildet:

<Wortklasse>	<Häufigkeit>	<Abstand>
...		
AFFAIRS	37	
MANAGER	37	
COMMUNITY	37	
OBRIEN	37	
O/BRIEN	17	(d = 1)
POLLY	23	
PO11Y	8	(d = 2)
PO1LY	5	(d = 1)
SCHWARTZ	15	
MIK0	15	
MIKO	3	(d = 1)
PAULA	8	

...

==========

Die Bildung von Repräsentanten kann je nach Anwendung mit weiterem Wissen unterstützt werden. So kann ein Wort entweder auf eine Zahl oder auf eine Alpha-Folge abgebildet werden, indem OCR-Ersetzungstabellen verwendet werden, die austauschbare Zeichenpaare definieren, wie 1 - L, 0 - O, 2 - Z, 6 - G, etc. Wenn darüberhinaus zu erlernenden Wortklassen Alternativenmengen bekannt sind - für Vornamen beispielsweise Spitznamen, wie Paula-Polly, Thomas-Tom, etc., kann auch diese Ersetzung vorgenommen werden. Beide Schritte können auf das Wörterbuch W1 51 angewendet werden, was zu einer weiteren Verschmelzung von Wortklassen führt.

Abschließend werden in den Erkennungsergebnissen alle Wörter, die im Wörterbuch W1 51 vorkommen, markiert und durch ihren Repräsentanten ergänzt. Diese Wörter werden im folgenden mit W1-Wörter bezeichnet.

An der Spitze vom Wörterbuch W1 51 stehen nun die häufigsten, bisher unbekannten Wortformen und die Wortklassen enthalten Schreibvarianten davon. So werden in der Anwendung der innerbetrieblichen Postverteilung bisher unbekannte Nach- und Vornamen und Teile von Abteilungsbezeichnungen im Wörterbuch W1 51 stehen. Darüberhinaus enthalten deren Wortklassen Schreibvarianten oder Varianten, die aufgrund der Eigenschaften des Lesesystems entstanden sind.

Ausgehend von den Repräsentanten der Wortklassen im Wörterbuch W1 51, die in den Erkennungsergebnissen als solche markiert sind, werden im nächsten Schritt nach FIG 4 Wortgruppen der Länge 2 bis n bestimmt, indem die Nachbarschaften von W1-Wörtern der Erkennungsergebnisse 62 untersucht werden. Für jedes W1-Wort wird dazu die rechte Nachbarschaft in einem Fenster der Breite k <= n durchsucht, ob darin weitere W1-Wörter sind. n-1 zunächst leere Wörterbücher werden in einem

Zwischenspeicher angelegt und Schritt für Schritt gefüllt. Ein n-Tupel wird dann in einen Wortgruppen-Zwischenspeicher aufgenommen 53, wenn n W1-Wörter gefunden worden sind und weniger als m weitere nicht W1-Wörter zwischen diesen n liegen. Wie beim Wörterbuch W1 51, wird auch hier die Auftretenshäufigkeit der einzelnen Wortgruppen der Länge n gespeichert.

Der Wahl der Werte von m und n hängt von der konkreten Anwendung ab. Für Werte n > 4 sind bei der Anwendung Adreßlesen keine signifikant häufigen Einträge mehr zu erwarten. m = 0 bedeutet, daß alle n W1-Wörter direkt aufeinanderfolgen. Gerade bei Paaren von Vornamen und Nachnamen kann jedoch ein zweiter Vorname hin und wieder die direkte Aufeinanderfolge unterbrechen, genauso wie Segmentierfehler der Lesemaschine vermeintliche Worthypothesen erzeugen können und damit eine direkte Aufeinanderfolge verhindern. Für die beschriebene Anwendung sind folglich m=1 und n=3 geeignete Werte.
In diesem Schritt werden folglich aus dem Wortgruppen-Zwischenspeicher n-1 Wörterbücher Wn 61 generiert, die häufige Wortsequenzen mit ihren Häufigkeiten für Paare, Triplets, etc. bis zu n-Tupel enthalten. In jedem Wörterbüch Wn 61 werden die Häufigkeiten der n-Tupel mit den Häufigkeiten der W1-Wörter der n-Tupel zu einer Maßzahl verrechnet. Jedes Wörterbuch Wn 61 wird nach absteigenden Maßzahlen sortiert, so daß wieder die signifikantesten Wortgruppen am Anfang eines jeden Wörterbuches Wn stehen 54.

Für obiges Beispiel sieht das Wörterbuch W2 folgendermaßen aus:

W2

COMMUNITY AFFAIRS	37
MANAGER COMMUNITY	37
POLLY OBRIEN	23
MIKO SCHWARTZ	15
PAULA OBRIEN	8

Das Wörterbuch W3 hat 3 Einträge, vorausgesetzt, daß der Name POLLY OBRIEN stets mit der Bezeichnung MANAGER COMMUNITY AFFAIRS kombiniert vorkommt und ein Zeilenumbruch in einem n-Tupel erlaubt ist.:

W3

MANAGER COMMUNITY AFFAIRS	37
POLLY OBRIEN MANAGER	23
OBRIEN MANAGER COMMUNITY	23

Wie beschrieben werden nun die Wortvorschläge der Wörterbücher Wn 61 (W2, W3, etc) entsprechend FIG 5 einem Operator zur Validierung vorgelegt. Durch Wissen über die zu erlernenden Worteinheiten 72 ist es an dieser Stelle möglich, Einträge in den Wörterbüchern W1, W2, .. Wn 51, 61 semantisch zu kategorisieren 71. So lassen sich in dieser Anwendung Einträge der semantischen Klasse <Name> zuordnen, indem in allgemeingültigen Vornamenslisten nachgeschlagen wird. Ähnliches gilt für die Semantikklasse <Abteilung>, die sich aus Schlüsselwörtern wie Department ableiten läßt.
Dieser Vorgang ist selbstverständlich auch automatisch ohne Operator durch Vergleich mit den Einträgen dieser Listen auszuführen.

Zu erfolgreich verteilten Sendungen sind die dazu erforderlichen Adreßelemente gefunden worden und sind als solche in den Erkennungsergebnissen gekennzeichnet. Wenn beispielsweise in der Anwendung der innerbetrieblichen Postverteilung Nachnamen und Vornamen erfolgreich gelesen worden sind, werden diese Ergebnisse in einer Statistik erfaßt; insbesondere wird die Häufigkeit der extrahierten Wörter, Paare, im allgemeinen von n-Tupeln, über definierte Zeitabschnitte td, z.B. für eine Woche, gespeichert, wobei die Sendungsart berücksichtigt werden kann. Als Ergebnis erhält man eine Verteilung der zu ex-

trahierenden Adreßelemente für eine Folge von Zeitabschnitten:

Zeitpunkt 1	
MELINDA DUCKSWORTH	123
ALFRED SCHMID	67
...	
Zeitpunkt 2	
MELINDA DUCKSWORTH	1
ALFRED SCHMID	85
...	
Zeitpunkt 3	
MELINDA DUCKSWORTH	2
ALFRED SCHMID	72
...	

Aus der so ermittelten Verteilung läßt sich ableiten, ob Wörterbucheinträge gelöscht werden sollen: Die Einträge werden in eine Liste zum Entfernen aus dem Wörterbuch eingefügt, wenn deren Häufigkeit sich von td_i zu td_{i+1} abrupt verringert und auf diesem Niveau in aufeinanderfolgenden Zeitabschnitten td_{i+k} bleibt (z.B. $k = 4$). So wird im obigen Beispiel die Person MELINDA DUCKSWORTH im Wörterbuch gelöscht. Dieser Ablauf kann zusätzlich auch über einen Bestätigungsvorgang geführt werden.

Patentansprüche

1. Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adreßlesen,
gekennzeichnet durch die Schritte:
- Zwischenspeicherung der vom OCR-Leser erzielten Leseergebnisse der Adressen einer vereinbarten Anzahl von Sendungsbildern oder innerhalb einer vereinbarten Zeitspanne gelesener Sendungsbilder, unterteilt in eindeutig gelesene Ergebnisse mit einer Übereinstimmung mit einem Wörterbucheintrag und in zurückgewiesene Leseergebnisse ohne Übereinstimmung mit einem Wörterbucheintrag,
- Bildung von Klassen von Wörtern mit dazugehörenden Repräsentanten oder zusammengehörenden Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse, bestehend jeweils aus n Adreßwörtern,n= 1,2,...a, mit den Wortabständen m, m= 0,1,...b, die bezogen auf jeweils einen bestimmten n- und m-Wert untereinander ein bestimmtes Ähnlichkeitsmaß nicht unterschreiten,
- Aufnahme mindestens der Repräsentanten derjenigen Klassen, deren Häufigkeit einen festgelegten Wert überschreiten, in das oder die Wörterbücher der zugeordneten Adreßbereiche.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß
- zur Klassenbildung eine Häufigkeitsliste aller vorkommenden Wörter oder Wortgruppen der zurückgewiesenen Leseergebnisse, nach deren Häufigkeit sortiert, erstellt wird,
- zu jedem Wort oder jeder Wortgruppe, beginnend mit dem häufigsten Wort oder der häufigsten Wortgruppe, das Ähnlichkeitsmaß mit allen übrigen Wörtern oder Wortgruppen bestimmt und in eine Ähnlichkeitsliste eingetragen wird,
- alle Wörter oder Wortgruppen in der Ähnlichkeitsliste mit einem Ähnlichkeitsmaß über einer festgelegten Schwelle dem aktuellen Wort oder der aktuellen Wortgruppe als Klasse zugeordnet werden,

- anschließend die Wörter oder Wortgruppen der jeweils gebildeten Klasse aus der Häufigkeitsliste entfernt werden.

3. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
der Repräsentant der jeweiligen Klasse von Wörtern oder Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse durch das/die kürzeste oder häufigste Wort oder Wortgruppe gebildet wird.

4. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
die zeitliche Häufigkeit der Wörter oder Wortgruppen der eindeutig gelesenen Adressen statistisch dahingehend ausgewertet werden, daß bei deren plötzlicher und über einen festgelegten Zeitraum andauernder Verringerung über eine festgelegte Schwelle die jeweiligen eingetragenen Wörter oder Wortgruppen aus dem Wörterbuch entfernt werden.

5. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
irrelevante Wörter der Leseergebnisse durch Vergleich mit in einer speziellen Datei gespeicherten Wörtern ermittelt und nicht in das Wörterbuch aufgenommen werden.

6. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
kurze Wörter ohne Abkürzungspunkt mit weniger als p Buchstaben nicht in das Wörterbuch aufgenommen werden.

7. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
in das Wörterbuch neben den Repräsentanten auch die Wörter und/oder Wortgruppen der dazugehörenden Klassen mit den Ähnlichkeitsmaßen und Häufigkeiten eingetragen werden.

8. Verfahren nach einem der Ansprüche 1 und 2, d a d u r c h g e k e n n z e i c h n e t, daß
für Wörtergruppen mit n Wörtern, n>1, wobei die Wörter untereinander einen Abstand von m Wörtern, m>=0, haben, ausgehend vom jeweiligen, für das Wörterbuch ermittelten Einzelwort die Adressen mit Fenstern der Breite von n+m Wörtern durchsucht werden und beim Finden von weiteren n-1 für das Wörterbuch ermittelten Einzelwörtern in den festgelegten Abständen m untereinander diese gefundenen Wortgruppen mit deren Häufigkeiten in das entsprechende Wörterbuch übernommen werden.

9. Verfahren nach einem der Ansprüche 1,2,7,8, d a d u r c h g e k e n n z e i c h n e t, daß
das Ähnlichkeitsmaß zwischen den Wörtern mit dem Levenshtein-Verfahren ermittelt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9, d a d u r c h g e k e n n z e i c h n e t, daß
die zu entfernenden Wörterbucheintragungen und die Neueintragungen ins Wörterbuch an einem Videocodierplatz angezeigt, kategorisiert und bestätigt werden.

11. Verfahren nach einem der Ansprüche 1 bis 9, d a d u r c h g e k e n n z e i c h n e t, daß
die ins Wörterbuch einzutragenden Wörter und/oder Wortgruppen vor deren Eintragung mit den Inhalten einer Datei verglichen werden, in der für die jeweilige Wörterbuchkategorie charakteristische, allgemeingültige Namen oder wenigstens Zeichenstrings gespeichert sind, und bei Übereinstimmung in das entsprechende Wörterbuch übertragen werden.

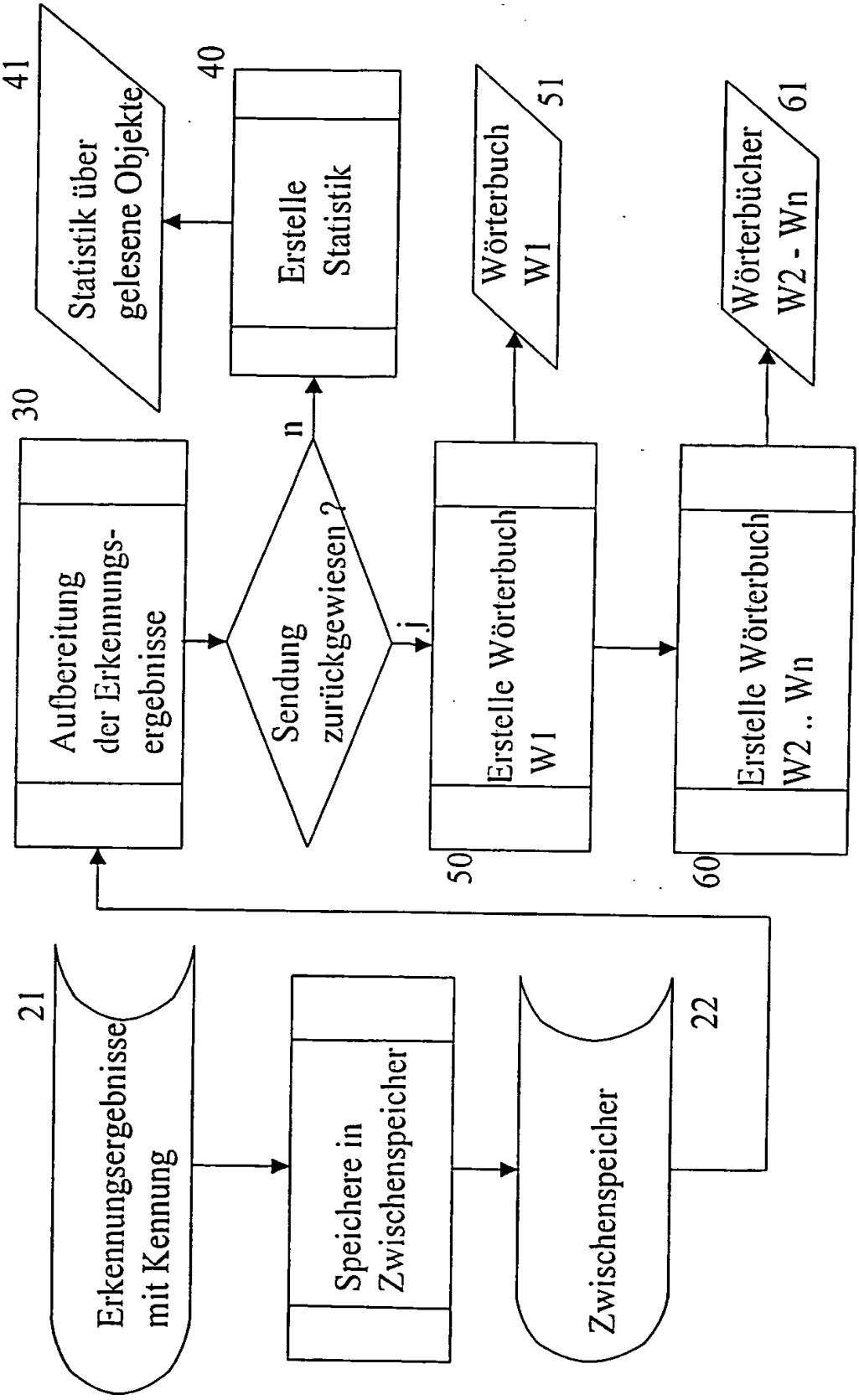

Erkennungsergebnisse mit Kennung
21
Speichere in Zwischenspeicher
Zwischenspeicher
22
Aufbereitung der Erkennungs-ergebnisse
30
Sendung zurückgewiesen ?
n
j
Erstelle Statistik
40
Statistik über gelesene Objekte
41
Erstelle Wörterbuch W1
50
Wörterbuch W1
51
Erstelle Wörterbuch W2 .. Wn
60
Wörterbücher W2 - Wn
61

FIG 1

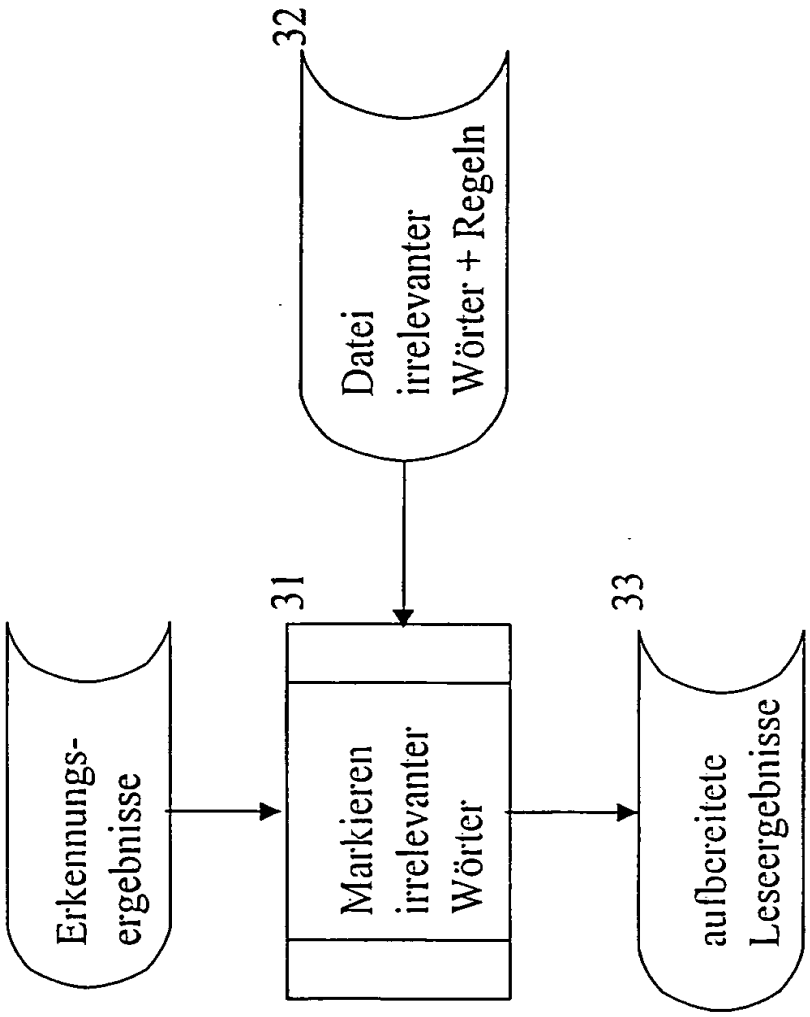
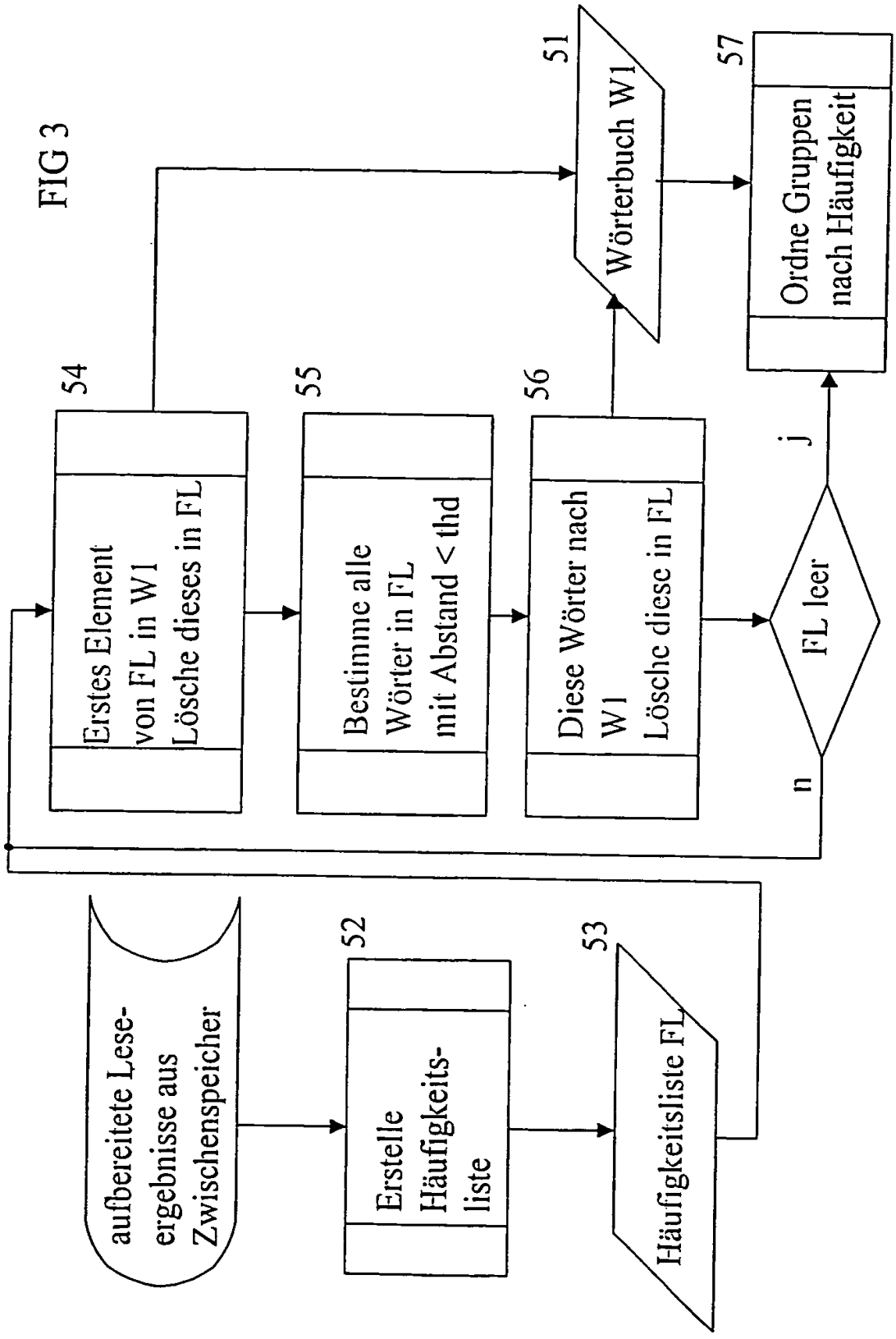

32
Datei
irrelevanter
Wörter + Regeln
31
33
Erkennungs-
ergebnisse
Markieren
irrelevanter
Wörter
aufbereitete
Leseergebnisse

FIG 2

FIG 3


aufbereitete Lese-ergebnisse aus Zwischenspeicher
52
Erstelle Häufigkeits-liste
53
Häufigkeitsliste FL
54
Erstes Element von FL in W1 Lösche dieses in FL
55
Bestimme alle Wörter in FL mit Abstand < thd
56
Diese Wörter nach W1 Lösche diese in FL
FL leer
n
j
51
Wörterbuch W1
57
Ordne Gruppen nach Häufigkeit

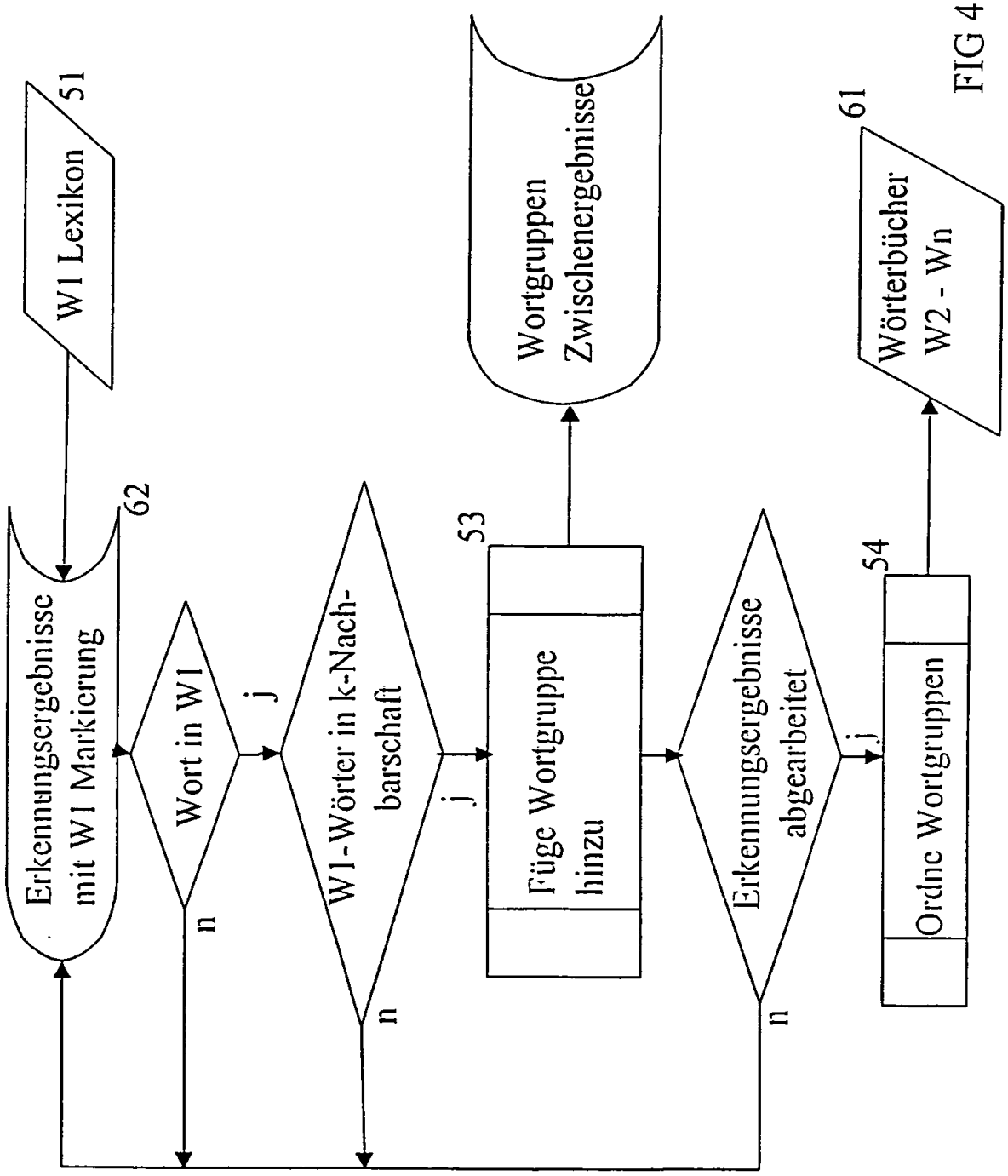

W1 Lexikon
51
Erkennungsergebnisse mit W1 Markierung
62
Wort in W1
j
n
W1-Wörter in k-Nachbarschaft
j
n
53
Füge Wortgruppe hinzu
Wortgruppen Zwischenergebnisse
Erkennungsergebnisse abgearbeitet
j
n
54
Ordne Wortgruppen
61
Wörterbücher W2 - Wn

FIG 4

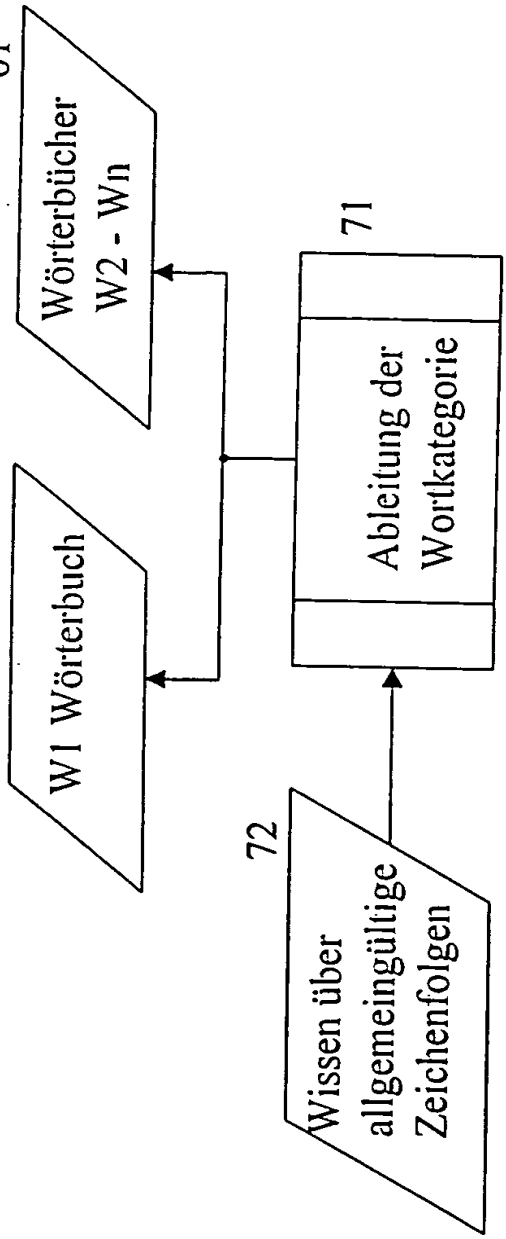

61
Wörterbücher
W2 - Wn
71
Ableitung der
Wortkategorie
51
W1 Wörterbuch
72
Wissen über
allgemeingültige
Zeichenfolgen

FIG 5

INTERNATIONAL SEARCH REPORT

Inter. onal Application No
CT/DE 00/01791

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/72 G06K9/62

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
INSPEC, WPI Data, IBM-TDB, EPO-Internal, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 754 671 A (SCHAEWE TIMOTHY J ET AL) 19 May 1998 (1998-05-19) column 11, line 49 -column 11, line 67; claim 1; figure 23 ---	1-11
A	JONATHAN J. HULL: "The Use of Global Context in Text Recognition" EIGTH INT. CONF. ON PATTERN RECOGNITION, 27 - 31 October 1986, pages 1218-1220, XP002153998 Paris, FRance page 1219, left-hand column, last paragraph; figure 2 ---	1-11
A	WO 95 15535 A (MOTOROLA INC) 8 June 1995 (1995-06-08) abstract ---	1-11

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 November 2000	18/12/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Granger, B

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/DE 00/01791

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	OKUDA T ET AL: "METHOD FOR THE CORRECTION OF GARBLED WORDS BASED ON THE LEVENSTEIN METRIC" IEEE TRANSACTIONS ON COMPUTERS,IEEE INC. NEW YORK,US, vol. C-25, no. 2, 1976, pages 172-178, XP000916729 ISSN: 0018-9340 abstract	9

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

[redacted]rmation on patent family members

[redacted]ten. ...nal Application No
PCT/DE 00/01791

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5754671 A	19-05-1998	US 5805710 A	08-09-1998
WO 9515535 A	08-06-1995	AU 669087 B	23-05-1996
		AU 1288895 A	19-06-1995
		BR 9405791 A	12-12-1995
		CA 2153684 A	08-06-1995
		CN 1117319 A,B	21-02-1996
		EP 0686291 A	13-12-1995
		JP 8506444 T	09-07-1996
		SG 46656 A	20-02-1998
		US 6005973 A	21-12-1999
		ZA 9409146 A	21-07-1995

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 00/01791

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G06K9/72 G06K9/62

Nach der internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06K

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
INSPEC, WPI Data, IBM-TDB, EPO-Internal, COMPENDEX

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	US 5 754 671 A (SCHAEWE TIMOTHY J ET AL) 19. Mai 1998 (1998-05-19) Spalte 11, Zeile 49 -Spalte 11, Zeile 67; Anspruch 1; Abbildung 23 ---	1-11
A	JONATHAN J. HULL: "The Use of Global Context in Text Recognition" EIGTH INT. CONF. ON PATTERN RECOGNITION, 27. - 31. Oktober 1986, Seiten 1218-1220, XP002153998 Paris, FRance Seite 1219, linke Spalte, letzter Absatz; Abbildung 2 ---	1-11
A	WO 95 15535 A (MOTOROLA INC) 8. Juni 1995 (1995-06-08) Zusammenfassung --- -/--	1-11

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
28. November 2000	18/12/2000

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Granger, B

2

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 00/01791

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	OKUDA T ET AL: "METHOD FOR THE CORRECTION OF GARBLED WORDS BASED ON THE LEVENSTEIN METRIC" IEEE TRANSACTIONS ON COMPUTERS,IEEE INC. NEW YORK,US, Bd. C-25, Nr. 2, 1976, Seiten 172-178, XP000916729 ISSN: 0018-9340 Zusammenfassung	9

Formblatt PCT/ISA/210 (Fortsetzung von Blatt 2) (Juli 1992)

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

...nales Aktenzeichen
PCT/DE 00/01791

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 5754671 A	19-05-1998	US 5805710 A	08-09-1998
WO 9515535 A	08-06-1995	AU 669087 B	23-05-1996
		AU 1288895 A	19-06-1995
		BR 9405791 A	12-12-1995
		CA 2153684 A	08-06-1995
		CN 1117319 A,B	21-02-1996
		EP 0686291 A	13-12-1995
		JP 8506444 T	09-07-1996
		SG 46656 A	20-02-1998
		US 6005973 A	21-12-1999
		ZA 9409146 A	21-07-1995

7 T

VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS

PCT

REC'D 23 JUL 2001
PCT

INTERNATIONALER VORLÄUFIGER PRÜFUNGSBERICHT

(Artikel 36 und Regel 70 PCT)

Aktenzeichen des Anmelders oder Anwalts	WEITERES VORGEHEN	siehe Mitteilung über die Übersendung des internationalen vorläufigen Prüfungsberichts (Formblatt PCT/IPEA/416)
1999 P02291WO		
Internationales Aktenzeichen	Internationales Anmeldedatum *(Tag/Monat/Jahr)*	Prioritätsdatum *(Tag/Monat/Tag)*
PCT/DE00/01791	31/05/2000	20/07/1999
Internationale Patentklassifikation (IPK) oder nationale Klassifikation und IPK G06K9/72		
Anmelder SIEMENS AKTIENGSELLSCHAFT et al		

1. Dieser internationale vorläufige Prüfungsbericht wurde von der mit der internationalen vorläufigen Prüfung beauftragten Behörde erstellt und wird dem Anmelder gemäß Artikel 36 übermittelt.

2. Dieser BERICHT umfaßt insgesamt 4 Blätter einschließlich dieses Deckblatts.

   ☒ Außerdem liegen dem Bericht ANLAGEN bei; dabei handelt es sich um Blätter mit Beschreibungen, Ansprüchen und/oder Zeichnungen, die geändert wurden und diesem Bericht zugrunde liegen, und/oder Blätter mit vor dieser Behörde vorgenommenen Berichtigungen (siehe Regel 70.16 und Abschnitt 607 der Verwaltungsrichtlinien zum PCT).

   Diese Anlagen umfassen insgesamt 1 Blätter.

3. Dieser Bericht enthält Angaben zu folgenden Punkten:

   - I ☒ Grundlage des Berichts
   - II ☐ Priorität
   - III ☐ Keine Erstellung eines Gutachtens über Neuheit, erfinderische Tätigkeit und gewerbliche Anwendbarkeit
   - IV ☐ Mangelnde Einheitlichkeit der Erfindung
   - V ☒ Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung
   - VI ☐ Bestimmte angeführte Unterlagen
   - VII ☐ Bestimmte Mängel der internationalen Anmeldung
   - VIII ☐ Bestimmte Bemerkungen zur internationalen Anmeldung

Datum der Einreichung des Antrags	Datum der Fertigstellung dieses Berichts
15/02/2001	19.07.2001
Name und Postanschrift der mit der internationalen vorläufigen Prüfung beauftragten Behörde: Europäisches Patentamt D-80298 München Tel. +49 89 2399 - 0 Tx: 523656 epmu d Fax: +49 89 2399 - 4465	Bevollmächtigter Bediensteter Kessler, C Tel. Nr. +49 89 2399 2582

INTERNATIONALER VORLÄUFIGER PRÜFUNGSBERICHT

Internationales Aktenzeichen PCT/DE00/01791

I. Grundlage des Berichts

1. Hinsichtlich der **Bestandteile** der internationalen Anmeldung (*Ersatzblätter, die dem Anmeldeamt auf eine Aufforderung nach Artikel 14 hin vorgelegt wurden, gelten im Rahmen dieses Berichts als "ursprünglich eingereicht" und sind ihm nicht beigefügt, weil sie keine Änderungen enthalten (Regeln 70.16 und 70.17)*):

Beschreibung, Seiten:

1-14	ursprüngliche Fassung				

Patentansprüche, Nr.:

2-11	ursprüngliche Fassung				
1	eingegangen am	05/07/2001	mit Schreiben vom	02/07/2001	

Zeichnungen, Blätter:

1-5	ursprüngliche Fassung				

2. Hinsichtlich der **Sprache**: Alle vorstehend genannten Bestandteile standen der Behörde in der Sprache, in der die internationale Anmeldung eingereicht worden ist, zur Verfügung oder wurden in dieser eingereicht, sofern unter diesem Punkt nichts anderes angegeben ist.

Die Bestandteile standen der Behörde in der Sprache: zur Verfügung bzw. wurden in dieser Sprache eingereicht; dabei handelt es sich um

- ☐ die Sprache der Übersetzung, die für die Zwecke der internationalen Recherche eingereicht worden ist (nach Regel 23.1(b)).
- ☐ die Veröffentlichungssprache der internationalen Anmeldung (nach Regel 48.3(b)).
- ☐ die Sprache der Übersetzung, die für die Zwecke der internationalen vorläufigen Prüfung eingereicht worden ist (nach Regel 55.2 und/oder 55.3).

3. Hinsichtlich der in der internationalen Anmeldung offenbarten **Nucleotid- und/oder Aminosäuresequenz** ist die internationale vorläufige Prüfung auf der Grundlage des Sequenzprotokolls durchgeführt worden, das:

- ☐ in der internationalen Anmeldung in schriftlicher Form enthalten ist.
- ☐ zusammen mit der internationalen Anmeldung in computerlesbarer Form eingereicht worden ist.
- ☐ bei der Behörde nachträglich in schriftlicher Form eingereicht worden ist.
- ☐ bei der Behörde nachträglich in computerlesbarer Form eingereicht worden ist.
- ☐ Die Erklärung, daß das nachträglich eingereichte schriftliche Sequenzprotokoll nicht über den Offenbarungsgehalt der internationalen Anmeldung im Anmeldezeitpunkt hinausgeht, wurde vorgelegt.
- ☐ Die Erklärung, daß die in computerlesbarer Form erfassten Informationen dem schriftlichen Sequenzprotokoll entsprechen, wurde vorgelegt.

4. Aufgrund der Änderungen sind folgende Unterlagen fortgefallen:

   ☐ Beschreibung, Seiten:
   ☐ Ansprüche, Nr.:
   ☐ Zeichnungen, Blatt: - - -

5. ☐ Dieser Bericht ist ohne Berücksichtigung (von einigen) der Änderungen erstellt worden, da diese aus den angegebenen Gründen nach Auffassung der Behörde über den Offenbarungsgehalt in der ursprünglich eingereichten Fassung hinausgehen (Regel 70.2(c)).

   (Auf Ersatzblätter, die solche Änderungen enthalten, ist unter Punkt 1 hinzuweisen;sie sind diesem Bericht beizufügen).

6. Etwaige zusätzliche Bemerkungen:

V. Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung

1. Feststellung

Neuheit (N)	Ja:	Ansprüche	1 - 11
	Nein:	Ansprüche	
Erfinderische Tätigkeit (ET)	Ja:	Ansprüche	1 - 11
	Nein:	Ansprüche	
Gewerbliche Anwendbarkeit (GA)	Ja:	Ansprüche	1 - 11
	Nein:	Ansprüche	

2. Unterlagen und Erklärungen
   siehe Beiblatt

Zu Punkt V

Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung

1. Gebiet: Schriftzeichenerkennung

2. Aufgabe: Bilden und/oder Aktualisieren von Wörterbüchern zum automatischen Adresslesen.

3. Nächstkommender Stand der Technik: Die WO-A-95/15535 offenbart die Aktualisierung eines Wörterbuches durch eine als "am wahrscheinlichsten" erkannte Buchstabenfolge entsprechend der Anweisung eines Benutzers.

4. Lösung: Zu zurückgewiesenen Wörtern oder Wortgruppen werden Klassen gebildet, und zumindest ein Klassenrepräsentant in das Wörterbuch aufgenommen, wenn die Häufigkeit des Vorkommens dieser Klasse einen Grenzwert überschreitet.

5. Erfinderische Tätigkeit: Diese Vorgehensweise wird von den verfügbaren Dokumenten nicht nahegelgt.

Patentansprüche

1. Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adresslesen,
g e k e n n z e i c h n e t d u r c h die Schritte:
- Zwischenspeicherung der vom OCR-Leser erzielten Leseergebnisse der Adressen einer vereinbarten Anzahl von Sendungsbildern oder innerhalb einer vereinbarten Zeitspanne gelesener Sendungsbilder, unterteilt in eindeutig gelesene Ergebnisse mit einer Übereinstimmung mit einem Wörterbucheintrag und in zurückgewiesene Leseergebnisse ohne Übereinstimmung mit einem Wörterbucheintrag,
- Bildung von Klassen von Wörtern oder zusammengehörenden Wortgruppen mit dazugehörenden Repräsentanten der zwischengespeicherten und zurückgewiesenen Leseergebnisse, wobei die Wortgruppen aus n Adresswörtern, n= 1,2,...a, bestehen, zwischen denen sich jeweils m, m = 0,1,...b, weitere Wörter befinden, und wobei die Wörter der Klassen von Wörtern oder die Wörter der Klassen von Wortgruppen, bezogen auf jeweils einen bestimmten n- und m-Wert, untereinander ein bestimmtes Ähnlichkeitsmaß nicht unterschreiten,
- Aufnahme mindestens der Repräsentanten derjenigen Klassen, deren Häufigkeit einen festgelegten Wert überschreiten, in das oder die Wörterbücher der zugeordneten Adressbereiche.

Translation

PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference 99P2291P	FOR FURTHER ACTION	See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/DE00/01791	International filing date (*day/month/year*) 31 May 2000 (31.05.00)	Priority date (*day/month/year*) 20 July 1999 (20.07.99)
International Patent Classification (IPC) or national classification and IPC G06K 9/72		
Applicant SIEMENS AKTIENGESELLSCHAFT		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 4 sheets, including this cover sheet.

   ☒ This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

   These annexes consist of a total of 1 sheets.

3. This report contains indications relating to the following items:

   - I ☒ Basis of the report
   - II ☐ Priority
   - III ☐ Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
   - IV ☐ Lack of unity of invention
   - V ☒ Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
   - VI ☐ Certain documents cited
   - VII ☐ Certain defects in the international application
   - VIII ☐ Certain observations on the international application

Date of submission of the demand	Date of completion of this report
15 February 2001 (15.02.01)	19 July 2001 (19.07.2001)
Name and mailing address of the IPEA/EP Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/DE00/01791

I. Basis of the report

1. With regard to the **elements** of the international application:*

☐ the international application as originally filed

☒ the description:
pages 1-14, as originally filed
pages _____, filed with the demand
pages _____, filed with the letter of _____

☒ the claims:
pages 2-11, as originally filed
pages _____, as amended (together with any statement under Article 19
pages _____, filed with the demand
pages 1, filed with the letter of 05 July 2001 (05.07.2001)

☒ the drawings:
pages 1-5, as originally filed
pages _____, filed with the demand
pages _____, filed with the letter of _____

☐ the sequence listing part of the description:
pages _____, as originally filed
pages _____, filed with the demand
pages _____, filed with the letter of _____

2. With regard to the **language**, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.
These elements were available or furnished to this Authority in the following language _____ which is:

☐ the language of a translation furnished for the purposes of international search (under Rule 23.1(b)).

☐ the language of publication of the international application (under Rule 48.3(b)).

☐ the language of the translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

☐ contained in the international application in written form.

☐ filed together with the international application in computer readable form.

☐ furnished subsequently to this Authority in written form.

☐ furnished subsequently to this Authority in computer readable form.

☐ The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

☐ The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. ☐ The amendments have resulted in the cancellation of:

☐ the description, pages _____

☐ the claims, Nos. _____

☐ the drawings, sheets/fig _____

5. ☐ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).**

* *Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rule 70.16 and 70.17).*

** *Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.*

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/DE 00/01791

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty (N)	Claims	1-11	YES
	Claims		NO
Inventive step (IS)	Claims	1-11	YES
	Claims		NO
Industrial applicability (IA)	Claims	1-11	YES
	Claims		NO

2. Citations and explanations

1. Field: graphic character recognition.

2. Problem of interest: creating and/or updating dictionaries for the automatic reading of addresses.

3. Closest prior art: WO-A-95/15535 discloses the updating of a dictionary using a character string recognised as being the "most likely" in line with user instructions.

4. Solution: classes are formed from rejected words or groups of words and at least one class representative is included in the dictionary when the frequency of occurrence of that class exceeds a certain limit.

5. Inventive step: the above procedure is not suggested by the available documents.

VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT
AUF DEM GEBIET DES PATENTWESENS

PCT

INTERNATIONALER RECHERCHENBERICHT

(Artikel 18 sowie Regeln 43 und 44 PCT)

Aktenzeichen des Anmelders oder Anwalts	**WEITERES VORGEHEN**	siehe Mitteilung über die Übermittlung des internationalen Recherchenberichts (Formblatt PCT/ISA/220) sowie, soweit zutreffend, nachstehender Punkt 5
99P2291P		
Internationales Aktenzeichen	Internationales Anmeldedatum *(Tag/Monat/Jahr)*	(Frühestes) Prioritätsdatum *(Tag/Monat/Jahr)*
PCT/DE 00/ 01791	31/05/2000	20/07/1999
Anmelder		
SIEMENS AKTIENGSELLSCHAFT		

Dieser internationale Recherchenbericht wurde von der Internationalen Recherchenbehörde erstellt und wird dem Anmelder gemäß Artikel 18 übermittelt. Eine Kopie wird dem Internationalen Büro übermittelt.

Dieser internationale Recherchenbericht umfaßt insgesamt 3 Blätter.

[X] Darüber hinaus liegt ihm jeweils eine Kopie der in diesem Bericht genannten Unterlagen zum Stand der Technik bei.

1. **Grundlage des Berichts**

   a. Hinsichtlich der **Sprache** ist die internationale Recherche auf der Grundlage der internationalen Anmeldung in der Sprache durchgeführt worden, in der sie eingereicht wurde, sofern unter diesem Punkt nichts anderes angegeben ist.

   [ ] Die internationale Recherche ist auf der Grundlage einer bei der Behörde eingereichten Übersetzung der internationalen Anmeldung (Regel 23.1 b)) durchgeführt worden.

   b. Hinsichtlich der in der internationalen Anmeldung offenbarten **Nucleotid- und/oder Aminosäuresequenz** ist die internationale Recherche auf der Grundlage des Sequenzprotokolls durchgeführt worden, das

   [ ] in der internationalen Anmeldung in Schriftlicher Form enthalten ist.

   [ ] zusammen mit der internationalen Anmeldung in computerlesbarer Form eingereicht worden ist.

   [ ] bei der Behörde nachträglich in schriftlicher Form eingereicht worden ist.

   [ ] bei der Behörde nachträglich in computerlesbarer Form eingereicht worden ist.

   [ ] Die Erklärung, daß das nachträglich eingereichte schriftliche Sequenzprotokoll nicht über den Offenbarungsgehalt der internationalen Anmeldung im Anmeldezeitpunkt hinausgeht, wurde vorgelegt.

   [ ] Die Erklärung, daß die in computerlesbarer Form erfaßten Informationen dem schriftlichen Sequenzprotokoll entsprechen, wurde vorgelegt.

2. [ ] **Bestimmte Ansprüche haben sich als nicht recherchierbar erwiesen** (siehe Feld I).

3. [ ] **Mangelnde Einheitlichkeit der Erfindung** (siehe Feld II).

4. Hinsichtlich der **Bezeichnung der Erfindung**

   [X] wird der vom Anmelder eingereichte Wortlaut genehmigt.

   [ ] wurde der Wortlaut von der Behörde wie folgt festgesetzt:

5. Hinsichtlich der **Zusammenfassung**

   [X] wird der vom Anmelder eingereichte Wortlaut genehmigt.

   [ ] wurde der Wortlaut nach Regel 38.2b) in der in Feld III angegebenen Fassung von der Behörde festgesetzt. Der Anmelder kann der Behörde innerhalb eines Monats nach dem Datum der Absendung dieses internationalen Recherchenberichts eine Stellungnahme vorlegen.

6. Folgende Abbildung der **Zeichnungen** ist mit der Zusammenfassung zu veröffentlichen: Abb. Nr. 1

   [X] wie vom Anmelder vorgeschlagen

   [ ] keine der Abb.

   [ ] weil der Anmelder selbst keine Abbildung vorgeschlagen hat.

   [ ] weil diese Abbildung die Erfindung besser kennzeichnet.

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 00/01791

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G06K9/72 G06K9/62

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06K

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
INSPEC, WPI Data, IBM-TDB, EPO-Internal, COMPENDEX

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	US 5 754 671 A (SCHAEWE TIMOTHY J ET AL) 19. Mai 1998 (1998-05-19) Spalte 11, Zeile 49 -Spalte 11, Zeile 67; Anspruch 1; Abbildung 23 ---	1-11
A	JONATHAN J. HULL: "The Use of Global Context in Text Recognition" EIGTH INT. CONF. ON PATTERN RECOGNITION, 27. - 31. Oktober 1986, Seiten 1218-1220, XP002153998 Paris, FRance Seite 1219, linke Spalte, letzter Absatz; Abbildung 2 ---	1-11
A	WO 95 15535 A (MOTOROLA INC) 8. Juni 1995 (1995-06-08) Zusammenfassung ---	1-11

-/--

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

° Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
28. November 2000	18/12/2000

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Granger, B

2

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	OKUDA T ET AL: "METHOD FOR THE CORRECTION OF GARBLED WORDS BASED ON THE LEVENSTEIN METRIC" IEEE TRANSACTIONS ON COMPUTERS,IEEE INC. NEW YORK,US, Bd. C-25, Nr. 2, 1976, Seiten 172-178, XP000916729 ISSN: 0018-9340 Zusammenfassung	9

INTERNATIONAL SEARCH REPORT

[redacted]mation on patent family members

International Application No
[redacted]CT/DE 00/01791

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5754671 A	19-05-1998	US 5805710 A	08-09-1998
WO 9515535 A	08-06-1995	AU 669087 B	23-05-1996
		AU 1288895 A	19-06-1995
		BR 9405791 A	12-12-1995
		CA 2153684 A	08-06-1995
		CN 1117319 A,B	21-02-1996
		EP 0686291 A	13-12-1995
		JP 8506444 T	09-07-1996
		SG 46656 A	20-02-1998
		US 6005973 A	21-12-1999
		ZA 9409146 A	21-07-1995